

Measuring User Understanding in Explainable Human-Robot Interaction: A Systematic Review

FERRAN GEBELLÍ*, PAL Robotics, Spain

PRADIP PRAMANICK*, University of Naples Federico II, Italy

TAMLIN LOVE*, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain

RAQUEL ROS, Artificial Intelligence Research Institute (IIIA-CSIC), Spain

ANAÍS GARRELL, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain

SILVIA ROSSI, University of Naples Federico II, Italy

ANTONIO ANDRIELLA, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain

GUILLEM ALENYÀ, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain

In explainable Human-Robot Interaction (HRI), the primary objective of explanations and transparent behaviour is often to enhance the user's understanding of the robot. Consequently, assessing understandability becomes essential for evaluating the effectiveness of such systems. However, the evaluation of explainable HRI lacks standardisation, with numerous competing measures of understandability currently in use. In this paper, we conduct a systematic review of the literature on the evaluation of explainable HRI, with a focus on how understandability is operationalised and measured. We identify 58 eligible papers that include user studies in which understandability is measured. We categorise these papers according to five main aspects: whether the measure is subjective or objective, whether or not understandability is treated as a multidimensional construct, whether the measure is quantitative or qualitative, the ecological validity of the study and the temporal aspect of the measurements. The results reveal some notable trends in explainable HRI, including the different objective, subjective, quantitative, and qualitative approaches to measuring understandability, as well as some clear gaps in the operationalisation of understandability. In particular, we identify a lack of in-the-wild studies and a limited number of measures that decompose understandability into multiple dimensions. Moreover, we find an absence of studies that assess understandability longitudinally. Based on our analysis of the reviewed literature, we offer a set of recommendations for researchers conducting user studies on explainability in HRI and highlight several open questions regarding the measurement of understandability.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Human-centered computing** → *HCI design and evaluation methods*; • **Computer systems organization** → External interfaces for robotics.

Additional Key Words and Phrases: Explainability, Transparency, Understandability, XAI, XHRI

*These authors contributed equally, and the author order has been randomly generated via <https://www.aeaweb.org/journals/policies/random-author%2Dorder/search?RandomAuthorsSearch%5Bsearch%5D=SHvdOB4nKVcP>

Authors' Contact Information: [Ferran Gebellí](mailto:ferran.gebelli@pal-robotics.com), ferran.gebelli@pal-robotics.com, PAL Robotics, Pujades, 77-79, 7-7, 08005 Barcelona, Spain; [Pradip Pramanick](mailto:pradip.pramanick@unina.it), pradip.pramanick@unina.it, University of Naples Federico II, Naples, Italy; [Tamlin Love](mailto:tlove@iri.upc.edu), tlove@iri.upc.edu, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Llorens i Artigas 4-6, 08028, Barcelona, Spain; [Raquel Ros](mailto:raquel.ros@iiia.csic.es), raquel.ros@iiia.csic.es, Artificial Intelligence Research Institute (IIIA-CSIC), Spain; [Anaís Garrell](mailto:anaïs.garrell@upc.edu), anaïs.garrell@upc.edu, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain; [Silvia Rossi](mailto:silvia.rossi@unina.it), silvia.rossi@unina.it, University of Naples Federico II, Italy; [Antonio Andriella](mailto:aandriella@iri.upc.edu), aandriella@iri.upc.edu, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain; [Guillem Alenya](mailto:guillem.alenya@iri.upc.edu), galenya@iri.upc.edu, Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

ACM Reference Format:

Ferran Gebellí, Pradip Pramanick, Tamlin Love, Raquel Ros, Anaís Garrell, Silvia Rossi, Antonio Andriella, and Guillem Alenyà. 2025. Measuring User Understanding in Explainable Human-Robot Interaction: A Systematic Review. 1, 1 (June 2025), 27 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Explainability is a critical requirement for embodied autonomous systems. With the rapid development of AI systems, there is a strong push from various stakeholders toward standardising the expected level of transparency in autonomous and intelligent systems. Notably, the IEEE P7001 standard on the transparency of autonomous systems defines explainability as “*the extent to which the internal state and decision-making processes of an autonomous system are accessible to non-expert stakeholders*” [126]. In the specific context of human-robot interaction (HRI), non-expert users who are unfamiliar with a robot’s internal mechanisms and decision-making process may find it difficult to understand the robot’s intent and abilities without explanations or other transparency mechanisms enabled by the robot. While users may differ in their prior knowledge and familiarity, our focus is on those not heavily involved in the design or development of autonomous robotic systems. Therefore, a measurement of the improvement of the knowledge that users have about a robot’s internal mechanisms is a strong indicator of the effectiveness of explanations [123, 128]. We refer to this level of the user’s knowledge about a robot’s decisions and behaviours that is facilitated by the explanatory information as *Understandability* (see Sec. 2.1.1 for a detailed discussion) [34, 70]. This user understanding serves as a mediating factor influencing key HRI measures such as usability, trust, performance in shared tasks, accountability, and ease of debugging [43, 63, 98, 109], as depicted in Fig. 1.

Unlike the typical applications of explainable AI (XAI), such as recommendation systems or singular classification models, explainability in HRI introduces unique challenges due to the physical and social presence of robots. Users tend to anthropomorphise robots, attributing intentions and agency to them, which leads to higher expectations for explanations, including explanations of physical movements and failures [75, 120]. In particular, autonomous robots can be subjected to a much wider variety of failures and unexpected behaviour, beyond what is typically considered in XAI settings, such as violating social norms, as well as timing, control, and actuator failures [15]. Apart from the necessity to explain broader categories of events, the robots’ embodiment further allows for new modalities to convey enriched multimodal explanations. For example, a robot’s physical limitations to reach certain poses can be expressed by motions [62] and verbal explanations can be complemented with graphics [37, 94], and gestures [50]. Further, distinct explanation types and communication methods (e.g., counterfactuals, verbosity, non-verbal cues) may lead to different cause attributions for similar robot behaviours [51, 72]. Therefore, while methods for evaluating understandability and previous empirical findings in XAI contexts are relevant, they cannot be trivially applied to HRI settings.

We next briefly describe the motivation and objective of this work, followed by its contributions to the increasingly active research area of explainable HRI.

Motivation and Objective. Despite efforts towards standardisation [126], measuring the effectiveness of explanations remains a challenge due to a lack of consensus on how explainability can be defined in terms of a measurable property of a system. The terms *transparency* and *explainability* are often used interchangeably [103, 126]. Previous works have also explored the relationships between explainability and other constructs such as *interpretability*, *predictability*, *legibility*, *understandability*, among others [43, 98, 109, 123]. Similarly, other literature reviews [43, 109, 123] make a distinction between evaluating the quality of explanations (factors such as accuracy and clarity) and the effects that

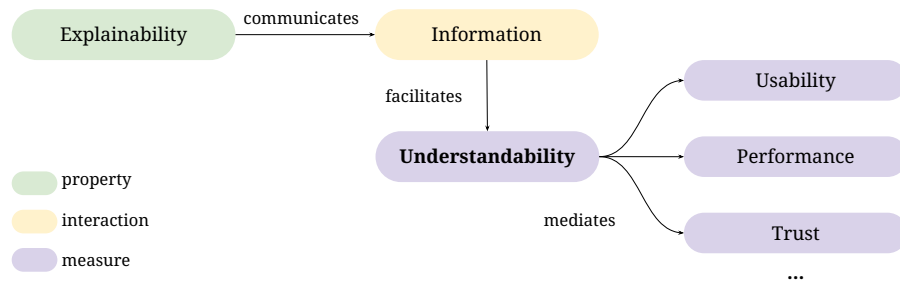


Fig. 1. Understandability as a mediating factor in fulfilling key desiderata in HRI—adapted and simplified from [43, 63, 98, 109].

explanations have on users. Langer et al. [63], Speith and Langer [109] suggest that the primary effect of explainability on users is a change in their understanding of the robot. Furthermore, human evaluation of the explainability of a system is considered both necessary and challenging [24], and this assessment relates to the ecological validity of the experiment (see Sec. 2.2.4).

Although research on explanation generation methods and their application in robotics is growing rapidly (Fig. 4), there is a lack of a comprehensive treatment of the concept of user understanding in prior works, particularly literature reviews and position papers. Relevant prior works include [80], which introduces the three phases of explanation, *generation*, *communication*, and *reception*, where explanation reception concerns “*how well the human understands the explanation*”. This categorisation forms the basis for a subsequent systematic review on explainable robotics that points out issues with user evaluation of explanations [2]. Other reviews on the topic focus on aspects such as the timing and interactiveness of explanations [4], implicit and explicit methods of conveying transparency [103], communication modalities [125], explanation generation methods [107], as well as on establishing definitive requirements for explanations in autonomous robots [100]. Although these reviews give a foundational overview and highlight important research challenges, the main point of discussion remains the methods of generating and communicating explanations in HRI, with much less emphasis on how explanations are received and understood by people. In contrast, our review focuses on evaluating how recent user studies on HRI capture the effectiveness of explanations in improving understandability.

Contributions. In this work, we address the following research question: *How do works in the field of explainable HRI conceptualise and operationalise understandability?* To that end, we make the following contributions:

- We devise a taxonomy (represented in Fig. 2) with which we can classify measures of understandability.
- Employing this taxonomy, we conduct a systematic review of user studies that evaluate understandability in explainable HRI, with a focus on the operationalisation and measurement of the construct.
- Based on the findings of our systematic review, we propose a set of recommendations for the evaluation of understandability in HRI, and identify open questions in the field.

In Sec. 2, we review the concept of understandability from the social sciences perspective. Then, we give an overview of the principles of HRI experiments that we use in our taxonomy to classify measures of understandability in our systematic review. Sec. 3 describes the search process, as well as the inclusion and exclusion criteria, following the PRISMA methodology [82]. Sec. 4 contains the analysis of the selected papers, starting with objective and subjective measures in Sec. 4.1. Sec. 4.2 highlights the prevalence of understanding as a unidimensional construct, discussing

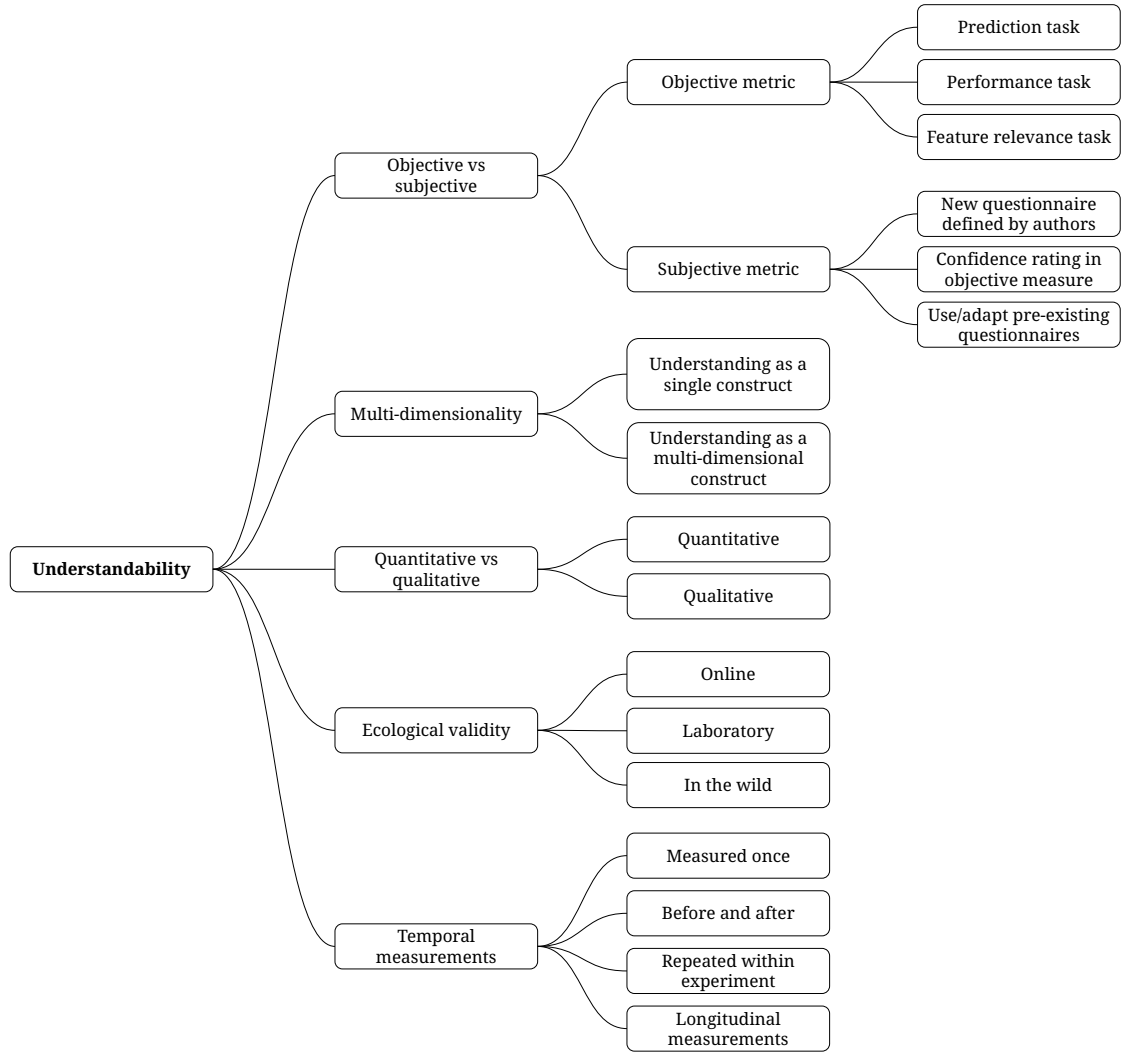


Fig. 2. Taxonomy of the literature review.

the few examples of multidimensional perspectives on understandability. Sec. 4.3 discusses the use of qualitative and quantitative measures of understandability. Then, Sec. 4.4 presents an overview of ecological validity, distinguishing online, in the lab and in-the-wild experiments. Sec. 4.5 presents an analysis of the frequency and timing of measurements. Based on this analysis, we propose a set of recommendations and open questions for conducting understandability evaluation in Sec. 5. Finally, we conclude the work in Sec. 6.

2 Background

In this section, we begin by reviewing how human understanding has been conceptualised in the social sciences. We then examine the key factors in the evaluation procedures of HRI user studies.

2.1 Understandability in the Social Sciences

This subsection presents an overview of the concept of human *understanding* from the point of view of the social sciences, and its relation to explanations.

The human understanding is a domain that has been widely studied in the social sciences. Understanding has historically been treated as a type of knowledge - the *knowledge of causes*. This idea is summarised by claiming that “understanding is not accomplished by acquiring some sort of superknowledge, but simply more knowledge” [70]. However, it has been argued that understanding can occur without relevant knowledge, and that knowledge can arise without related understanding [87]. Following this view, understanding would be a type of cognitive achievement. To address those critiques, authors in [34] propose an adapted version of the classic view, which considers understanding as *knowledge of causes*, while accounting for the cases where understanding can arise from non-causal sources, or from very brief knowledge.

Regarding the perspective from the social sciences on the relationship between explanations and understanding, explanations have been considered as an attempt to communicate information to others, so recipients of explanations can expand their understanding [49]. In turn, trying to explain to another, or even to oneself, often makes the explainers aware of the incompleteness of their own understanding [99], which in turn can help improve that understanding. Although humans often overestimate their level of understanding [48], they rarely believe to have a complete understanding. Humans deal with recognised gaps of understanding by distorting our beliefs or by outsourcing the gaps to the knowledge that other people have. In this way, humans stop upon reaching a “working understanding” and manage to get by with highly incomplete or partial explanations [49].

In summary, we observe from the social sciences research that: (1) human understanding is tied to knowledge, although it is not clear how tightly coupled they are; (2) explanations are a mechanism to transmit but also refine understanding; and (3) humans usually have an incomplete and evolving self-perception of understanding. These considerations are key to surveying how understanding can be measured in HRI, as we explore in the following subsections.

2.2 Evaluation Procedures in HRI User Studies

In this subsection, we explore different evaluation procedures in HRI user studies, intending to provide the key background concepts for the taxonomy from Fig. 2 later defined in Sec. 3.2.

2.2.1 Objective and Subjective Measures. In HRI user studies, measurements are commonly categorised as either objective or subjective [3, 42]. Subjective measures refer to ‘self-reported attitudes, thoughts, emotions, and moods of participants, collected through participants’ verbal responses,’ whereas objective measures are defined as ‘behavioural indicators that can be measured independently of participants’ stated opinions’ [42]. Most constructs can be assessed using either type of measure, and employing a combination of both is generally recommended [42].

Understandability measurements have also been classified into objective and subjective [95]. Objective understandability is the actual comprehension of the system [95], usually measured as the accuracy of the user’s mental model of the system in a proxy task [43]. A key aspect in evaluating objective understandability is choosing an appropriate proxy task because the selected task should “maintain the essence of the target application” [24]. One of the most widely recognised proxy tasks is *forward simulation* [24, 64, 71], which involves requiring participants to predict what the robot would do in hypothetical situations. This can be achieved through different methods, such as explicit prediction tasks, but also indirectly through think-aloud methodologies or structured interviews [43]. Subjective understandability is the

user-perceived and self-rated level of understanding, and is usually measured through questionnaires [95]. Subjective understandability has been inaccurately utilised as an indirect way to measure objective understandability [109], as research in social sciences shows that people tend to have a wrong perception of their understanding [117]. Some user studies have supported the idea that the subjective level of understandability is initially high but gradually declines as time progresses [58, 99], while other works [59] report the opposite, i.e., that subjective understandability begins at a lower level and then increases over time.

2.2.2 Understandability as a Multi-dimensional Construct. Works in the social sciences have suggested different types of understanding [5, 6, 18]. *Objectual understanding* relates to situations where *S understands X*, being *S* the subject and *X* a person, system, or language. *Propositional understanding* occurs when *S understands that something is the case*, while *interrogative understanding* refers to the circumstances when *S understands what/how/why is the case*. *Explanatory understanding* is the most important form of *interrogative understanding*, and is dedicated to the *why* questions. The different types are not completely independent, but they hold unclear relations to each other [5].

Similarly, Bloom's taxonomy [56] proposes a set of cognitive dimensions (*Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation*) which form a taxonomy of what is expected or intended for students to learn. A later work [10] relates this taxonomy to explanations using the theory of processes from Basic Formal Ontologies (BFO) [106]. In a first attempt to define distinct dimensions of understanding in XAI, Bloom's taxonomy is employed as a theoretical heuristic for the measurement of the dimensions of understandability. These dimensions could potentially be visualised over time with separate curves for each of them.

Apart from dimensions or types, understandability is contended to come in different degrees, which vary in depth and breadth [6]. To define these degrees, there are 3 possible approaches: define the minimal understanding, the maximal understanding, or determine what it means to understand something to a certain degree [6].

These theoretical categorisations of understandability indicate the presence of different dimensions that can potentially be measured separately. However, up to the authors' knowledge, no prior works have provided a more practical categorisation of the multidimensionality in understanding applicable to the field of explainable HRI.

2.2.3 Quantitative and Qualitative measures. Quantitative measurements (e.g. questionnaires) involve numerical data and typically require statistically significant sample sizes to support generalizable conclusions, while qualitative (e.g. semi-structured interviews, focus groups) approaches assume that the phenomena under study are context-dependent and not easily reduced to discrete variables, thereby necessitating interpretive analysis by the researchers [122].

There are several challenges in qualitative evaluations. The reproducibility of qualitative measurements can be low, especially in some study settings, such as ethnographic research or participatory design, which can be mitigated by providing more transparency on procedures, methods, and reported interactions with participants [36]. However, a literature review on qualitative evaluations in HRI concludes that there is a high variance in the rigour with which the approaches are applied [122]. Nonetheless, while qualitative measurements may not possess the precision of hypothesis-driven experimental studies, they can still systematically, rigorously, and formally capture holistic, multifactorial, and emergent data [104]. Complementary approaches and methodologies are needed to increase robustness, as defining a standardised approach is complex [23].

Quantitative evaluations are favoured by a tendency in HRI to aim for precision, characterised by clearly defined hypotheses [122] that can later be proved to be statistically significant. However, quantitative evaluations present challenges as well. Real behaviour data has been reported to differ from reported data in surveys due to the Hawthorne effect

[92], which refers to a “change in the behaviour of people because they feel observed” [8]. This affects both qualitative and quantitative assessments, but for qualitative assessments, observations can be designed to be less intrusive.

Quantitative and qualitative measures are complementary rather than mutually exclusive, leading to recommendations that both be considered in HRI research [3]. A common approach involves using qualitative methods to develop an initial theoretical framework, which is then tested through quantitative methods. However, the reverse sequence can also be valuable—for example, using qualitative analysis to interpret unexpected quantitative findings [104]. Furthermore, the boundary between qualitative and quantitative methods is becoming increasingly blurred, as traditionally qualitative techniques such as textual analysis are now often supported by computational tools for quantitative analysis [42].

2.2.4 Ecological Validity. In the HRI field, three levels of ecological validity have been identified: laboratory, online, and field studies in the wild [42].

Online studies are carried out through “crowdsourcing” platforms and are becoming increasingly employed [42] since they allow capturing data from many participants from a relatively controlled demographic. This enables obtaining statistical significances due to a large sample size in a relatively short time, since the deployment of a robot is not necessary, or the robot can be highly teleoperated. Participants do not interact with real robots, but receive pictures or videos of robots and are then asked to respond to some questionnaires. However, participants may not be committed to the study [135], and more importantly, the validity of online studies is low when real interactions with robots are needed to provide genuine responses to in-person interactions [13].

In other cases, users interact with real robots in laboratory settings. Normally, in those cases, the robot’s tasks are simplified, and the participants are given previous information to put them into context. The controlled environment still allows for the establishment of causal relationships and enables strict replication of conditions, but the population sample is often biased — typically pooled from university students— and the conclusions might not be generalisable to the real world outside of the lab [42].

Although online and laboratory studies can provide meaningful findings [43], it is not guaranteed that they could be replicated in an in-the-wild setting [8], where users can engage freely in an application-relevant environment. Even though the difference between a laboratory study and an in-the-wild study is not always black and white, since there are degrees “wilderness” [101], it has been argued that “we have a limited understanding of how people will respond to robots in complex social settings and how robots will affect social dynamics in situ” [47]. Despite the extensive advocates for in-the-wild studies [8, 47, 83, 92], a review study revealed that three-quarters of HRI user studies are lab studies [7]. The main reasons are the challenges posed by in-the-wild studies, as technology and resources limit the feasibility of these studies [8], and often there is a need to deal with multiple participants interacting at once, as groups engage more than individuals with robots in the wild [86].

More specifically to XAI, a taxonomy of evaluation [24] makes a distinction between functionally-, human- and application-grounded evaluations. In functionally-grounded evaluations, there are no humans involved, and proxy tasks are used. This is the first stage to test the actual generation of explanations (e.g., in terms of faithfulness and content quality), but not yet any of its effects on users. Then, in the human-grounded evaluation, tasks are kept simple, but humans are involved. This would include online and laboratory studies. Finally, in application-grounded evaluations, real humans evaluate systems with real-world tasks.

2.2.5 Temporal Measurements. Not only for understandability measurements but generally in HRI, participants interact only once with the robot [7], typically following the same procedure: (1) there is a short briefing session, (2) participants answer demographic and robot-prior-knowledge questions, (3) participants engage with the robot and (4) a post-session

is used to collect the relevant dependent variables. In some other studies, the relevant measures from the post-session are also included in the pre-session, allowing the measurement of changes due to the interaction with the robot. Finally, in other works, participants interact multiple times over time with the robot, with multiple sessions to collect measurements and study their evolution.

In many cases, users are not used to interacting with robots. However, the novelty effect is typically treated as noise that needs to be reduced instead of a valuable source of information, which should be considered together with user expectations and attributed anthropomorphism [105]. It has been proposed that studies should give greater attention to novelty effects, especially considering how curiosity will influence engagements when the novelty effect is still strong [92]. HRI is concerned with day-to-day interactions, and it is necessary to study how the user's behaviour will evolve once the novelty effect wears off [8]. In fact, the time needed to revoke the novelty effect has been proposed as the main constituent of a long-term interaction [65], while a minimum of three sessions across three consecutive days [73] offers a more functional definition.

Despite recommendations to account for the novelty effect and to conduct longitudinal experiments due to numerous advantages, such as achieving unique interactions or being able to explore the long-term robot adoption [73], a survey found that only 5 out of 96 HRI studies involved more than a single interaction [7]. Longitudinal studies pose several practical challenges, including identifying suitable, realistic and useful use cases, recruiting and retaining participants, determining interaction frequency, and securing the necessary resources [65, 73]. These challenges are further compounded by the impossibility of applying Wizard-of-Oz setups when robots are expected to operate autonomously over extended periods [61, 73]. Additionally, users often expect some degree of personalisation [61, 65, 73], which introduces ethical and privacy concerns, such as the need for facial recognition and the storage of personal interaction data [46]. Other ethical considerations include the potential formation of affective bonds with robots and the safety of participants in unsupervised, autonomous scenarios [65].

3 Systematic Literature Review

To better understand how understandability is operationalised in explainable HRI, we conducted a systematic review of the field using the PRISMA methodology [82], depicted as a flowchart in Fig. 3. This consisted of an identification and screening process (Sec. 3.1), after which the eligible papers were subjected to classification using our proposed taxonomy of understandability measures (Sec. 3.2).

3.1 Identification and Screening Process

To cast a wide net for relevant papers during the identification process, we conducted our search over three popular databases - the ACM Digital Library¹, IEEE Xplore² and Scopus³. To focus on the particular topic of understandability in explainable HRI, we filtered the search by a number of search terms. Firstly, “*robot** or *human-robot interaction* or *human robot interaction* or *hri*” was used to narrow down the search to the field of HRI. Next, “*explainability* or *explanation* or *transparency* or *interpretability* or *legibility*” was used to retrieve papers on transparency in HRI. The search term “*understanding* or *understandability*” was used to limit the search to papers that discuss the concept of understandability in some way. And finally, “*user study* or *experiment* or *survey*” was used to retrieve papers that actually operationalise transparency and understandability in experiments. Furthermore, to focus on recent research,

¹<https://dl.acm.org/>

²<https://ieeexplore.ieee.org/Xplore/home.jsp>

³<https://www.scopus.com>

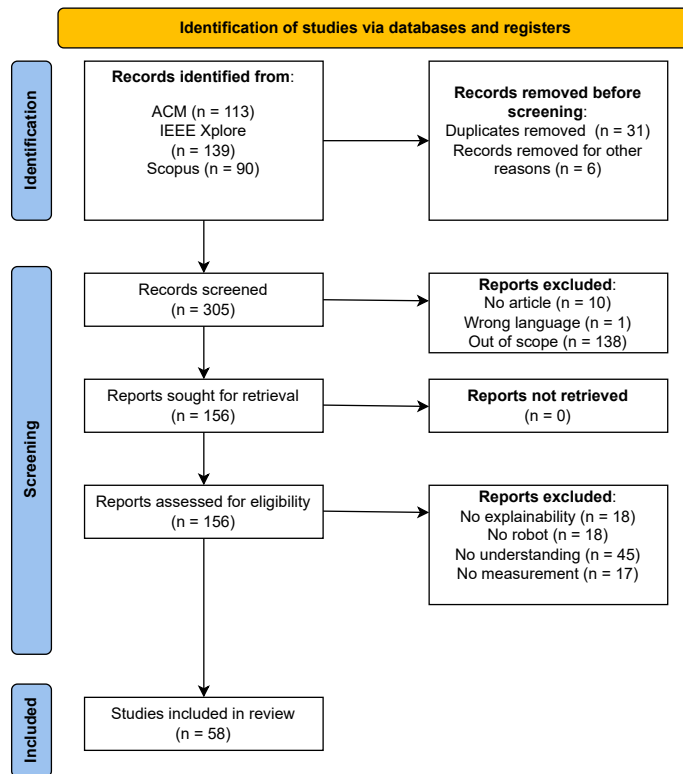


Fig. 3. PRISMA flowchart depicting the process of selecting the eligible articles for further analysis.

we limited our search to the period between January 1 2019 and December 2 2024, which is the date the search was conducted. With these filters in place, our search yielded 342 results. Of these, 31 papers were identified as duplicates and removed automatically, and a further 6 papers labelled as “conference reviews” were deemed unsuitable due to the type of document, leaving 305 papers to be screened.

Following the PRISMA methodology, we conducted several stages of screening to arrive at a set of eligible papers. Each stage was carried out by three researchers using an agreed-upon set of criteria. Discussions were held before, during, and after each stage to refine the criteria and handle ambiguous cases. In the first stage, each paper was screened (by assessing the title, abstract and keywords) to determine whether or not it should be retrieved for further screening. Of the 305 papers, 10 papers were excluded as they were deemed not to be scientific articles, 1 paper was excluded for not being in English, and 138 papers were determined to be clearly outside the scope of this review. Following this stage, 156 papers were sought for retrieval and full versions of all 156 were successfully retrieved.

In the final stage of screening, we removed ineligible papers with four exclusion criteria. First of all, papers needed to involve some sort of explanation or transparency. Papers without any (explicit or implicit) communication of information from a robot to a human were excluded (*no explainability*). 18 papers were excluded for this reason. Secondly, papers were excluded if they did not involve a robot, real or simulated (*No robot*). Papers with other kinds of agents (such as reinforcement learning agents) were included on a case-by-case basis, determined by their applicability to robotics domains. A further 18 papers were excluded for this reason. Next, we excluded any papers that did not at least mention the concept

of understandability (*no understanding*). Eligible papers were required to link the explainability component of their work to a person’s understanding of some aspect of the robot. 45 papers were excluded for this reason. Papers that mentioned understandability, but did not measure it in some way, were also excluded (*no measurement*). 17 papers were excluded for this reason. Note that we included papers that did not directly measure understandability but did employ proxy measures linked to users’ understanding of certain aspects of the robot (see our discussion of performance measures in Sec. 4.1).

At the end of the screening process, 58 eligible papers had been identified and are discussed in greater detail in the following sections.

3.2 Taxonomy of Understandability Measures

In order to further discuss and compare the eligible papers identified through the screening process, we devise a taxonomy with which we can classify the understandability measures employed in each work, depicted in Fig. 2. The categories we use are:

- **Objective vs. Subjective** - classifying whether a given measure is objective or subjective, as discussed in Sec. 2.2.1. During the review, we identified several sub-categories, which are discussed in greater detail in Sec. 4.1.
- **Multi-dimensionality** - classifying whether understandability is considered as a single construct or as a collection of related constructs by a given work, following our discussion in Sec. 2.2.2.
- **Quantitative vs. Qualitative** - classifying whether a given measure is quantitative or qualitative, as discussed in Sec. 2.2.3.
- **Ecological Validity** - classifying whether understandability is measured in an online, laboratory or in-the-wild setting, as discussed in Sec. 2.2.4.
- **Temporal Measurements** - classifying whether understandability is measured either only once during an experiment, before and after an explanatory intervention, repeated multiple times in a single experiment session, or repeated across multiple sessions in a longitudinal fashion, as discussed in Sec. 2.2.5.

During our categorisation, each paper received a classification in each category. In the cases where a paper used multiple measures of understandability or conducted multiple experiments, it would receive multiple categories. As with the screening process, the 58 papers were divided among three researchers for classification. In the case of doubt or ambiguity, papers were discussed and classified collectively. The results of this classification process are discussed in the following section.

4 Analysis

Before discussing how the reviewed papers fall into each of the five categories in our taxonomy, we begin by discussing key statistical trends that emerged from our review.

Our first observation is that, despite the filters discussed in Sec. 3, the papers are diverse in both subject area and publication venue. The eligible papers span a variety of target domains, including robot navigation [41, 74, 116], socially assistive robots [85, 134], manufacturing and other industrial applications [67, 90], autonomous vehicles [81], search and rescue [19], and various drone applications [1, 40, 132]. The eligible papers have been published in a variety of venues, with the most popular being RO-MAN⁴ (10 papers), HRI⁵ (10 papers, including late-breaking reports), THRI⁶

⁴IEEE International Conference on Robot and Human Interactive Communication

⁵ACM/IEEE International Conference on Human-Robot Interaction

⁶ACM Transactions on Human-Robot Interaction

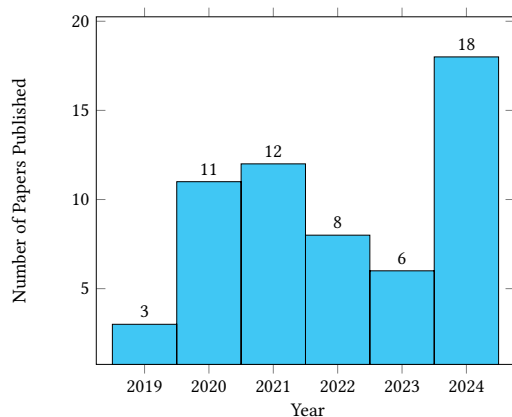


Fig. 4. The number of eligible papers by year of publication.

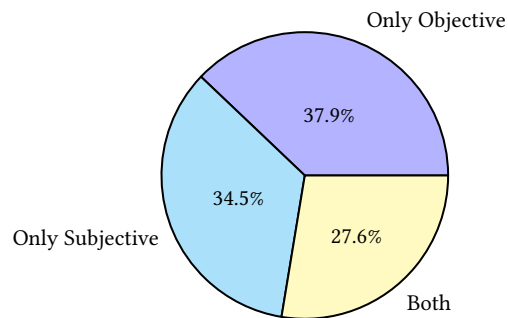


Fig. 5. Categorisation by the use of objective and subjective measures.

(6 papers) and IROS⁷ (4 papers). 41 eligible papers (~71%) appear in the proceedings and companion proceedings of conferences, while 17 (~29%) are published in journals.

Of the 58 eligible papers, just under a third (18, ~31%) were published in 2024 alone, indicating a growth in popularity for user studies in explainable HRI that measure understandability in some way (see Fig. 4).

4.1 Objective and Subjective Measures

Beginning with the first dimension of our taxonomy, we categorise each of the eligible papers based on their use of either objective or subjective measures, as defined in Sec. 2.2.1. The categorisation of each paper is shown in Table 1. Of the 58 eligible papers, 22 (~38%) make use of only objective measures, 20 (~34%) make use of only subjective measures, and 16 (~28%) make use of both. This indicates a fairly even split between the two types of measures, with a significant percentage employing both, as indicated in Fig. 5.

With a focus on objective measures, we observe a notable lack of standardised approaches to assessing understandability. Instead, such measures are often highly specific to the domain and the particular robotic task. Nevertheless, broad commonalities can be identified. In particular, we distinguish three main categories of objective measures: *predictive measures*, *feature relevance measures*, and *performance measures* (see Fig. 6).

Predictive measures operationalise understandability by asking a user to predict some aspect of the robot based on some context, in the vein of forward simulation [24]. For example, in [30], participants are asked to predict what object a robot arm is going to grasp. In [55], participants are asked to predict how the robot would resolve an ethical problem. In [40], participants are asked to predict the type and intensity of emotion expressed by a drone based on facial expressions rendered on an attached display. Of the 38 papers using objective measures, 25 (~66%) use at least one predictive measure.

In contrast, *feature relevance measures* operationalise understandability by asking users to identify important features in the robot's state that lead to some particular behaviour or decision. For example, in [102], participants are given a list of features used by a robot and asked to select the ones that are relevant to its decision-making in one of two toy domains. In [28], participants are asked to select reasons (i.e. features of the environment) that explain why a robot's

⁷IEEE/RSJ International Conference on Intelligent Robots and Systems

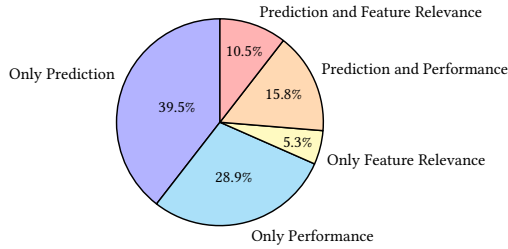


Fig. 6. Categorisation of objective measures.

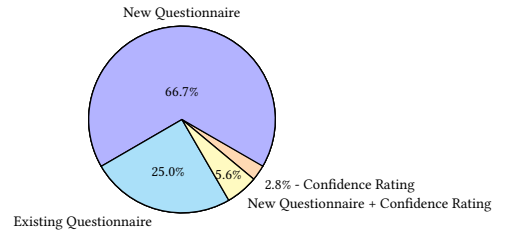


Fig. 7. Categorisation by subjective measures.

Objective Measure	Papers
Prediction-based	[1], [12], [16], [17], [20], [26], [30], [40], [44], [55], [66], [67], [72], [78], [81], [84], [89], [90], [102], [108], [113], [130], [132], [133], [134]
Performance-based	[1], [16], [17], [19], [21], [31], [35], [60], [74], [84], [88], [96], [98], [108], [114], [131], [132]
Feature relevance-based	[28], [81], [91], [102], [113], [130]
Subjective Measure	Papers
New questionnaire defined by authors	[1], [16], [19], [22], [39], [40], [41], [45], [66], [67], [79], [78], [85], [90], [96], [97], [98], [110], [111], [113], [115], [116], [121], [124], [133], [134]
Confidence rating in objective measure	[30], [66], [67]
Use/adapt pre-existing questionnaires	[9], [27], [28], [38], [75], [76], [81], [93], [127]

Table 1. Categorisation of objective and subjective measures. Papers appearing in multiple rows use multiple measures.

navigation path is not optimal. In [130], participants are asked to rate the importance of specific factors in the value function used by a team of robot scouts. Overall, only 6 papers (~16%) employ a feature relevance measure.

Finally, *performance measures* do not directly operationalise understandability, but instead use proxy measures, typically domain-dependent, to measure how well a user performs a task, with the assumption that improved understanding correlates with improved task performance. These measures are especially useful when the human and the robot must engage in some collaborative task. For example, in [21], the authors measure whether participants understand the intent of a robot in a social navigation scenario by assessing the behaviour of the human (either crossing in front of the robot or following behind). In [114], participants must determine whether a robot's navigation plan is optimal or not. In [31], participants must determine the errors made by a robot scanning objects in a room. 17 papers (~45%) make use of a performance measure in some way.

Turning to the subjective measures, we categorise each of the 36 papers using subjective measures based on the source of the measure employed in the paper, in order to identify measures that are commonly used by the explainable HRI community (see Fig. 7). However, we note that the majority of papers (26, ~72%) define their own measures, typically consisting of Likert scale questionnaires (e.g., [19, 41, 111]) or open-ended questions to extract qualitative data (e.g., [115, 121, 124]). Apart from these works, 3 papers (~8%) require users to rate their confidence in their objective understandability answer [30]. The remaining 9 papers (25%) directly use or adapt existing scales to subjectively measure understandability. Interestingly, none of these papers use the same scales, each employing a different questionnaire from the literature (taken from [11, 25, 33, 43, 54, 57, 68, 69, 119]). Overall, these results indicate an absence of standardisation when it comes to subjectively measuring understandability. However, many of the subjective questionnaires employ similar questions with different phrasings. Many questions ask participants to rate the extent to which the robot is

understandable, predictable, intentional, etc. (for example, compare “Pepper’s behaviour was understandable” [110], “The robot moved as I expected” [41], and “The robot motion clearly conveys its target to me” [133]). Others focus on the explanation, asking participants to rate how understandable or clear they find the explanations themselves (for example, compare “I have understood the information presented to me by the robot” [39], “The robot’s communication to me was clear” [41], and “Pepper’s explanation was understandable” [110]). This indicates that, while standardisation is low when it comes to subjective measures in the explainable HRI community, there are common needs that could form the basis of new standards in subjectively evaluating understandability.

Among the 36 papers utilising subjective measures, 26 (72%) employ Likert scales. 5-point and 7-point Likert scales are the most common, with 14 and 11 papers using them, respectively, and one paper using both [116]. Additionally, one paper uses a 6-point Likert scale [110], and another utilises a 9-point scale [127]. In contrast, six papers (17%) use open-ended questions. Other variations include a 1-to-7 scale from “never” to “always” [98], categorical “Yes/No/Maybe” responses [85], a confidence rating from 0 to 10 [30], two papers using a 1-to-7 confidence scale (from “Very Unsure” to “Very Confident”) [66, 67], and one paper using a multiple-choice scale to rate “clearness” with the options “It was clear”, “It was not clear”, and “Not sure” [134].

As stated before, 16 (~28%) papers make use of both objective and subjective measures. Some of these papers explicitly distinguish between *objective understandability* (measured objectively) and *perceived understandability* (measured subjectively). For example, in [28], subjective *perceived understandability* is distinguished from objective *map understanding* (a feature relevance measure). In [113], subjective *perceived understanding* is distinguished from objective *policy understanding* (a prediction measure). In [66, 67, 67], confidence ratings on the objective measures are used as subjective understandability measures.

4.2 Understandability as a Multidimensional Construct

The vast majority of eligible papers (55, ~95%, Fig. 8) consider understandability as a single, unidimensional construct and operationalise it accordingly. Even though several studies did incorporate multiple measures of understandability, the primary objective was to improve the robustness of the measurements, rather than measuring different dimensions or types of understandability as introduced in Sec. 2.2.2. For example, simulating a rover operation task, the authors in [113] employ a subjective measure of *perceived understanding* (7-point Likert, “I understand the behavior of the rover”) as well as two objective measures - a prediction measure (selecting the correct action for a given state) and a feature relevance measure (“Which of the parameters mattered the LEAST for the rover to choose an action?”). Similarly, considering decisions made by swarms of small unmanned drones, the study in [1] uses both a quantitative performance measure (“human-agent partnership” measured by observing participant interaction rates with video stream state information) and a qualitative analysis of participants’ open-ended responses (categorising challenges the participants identified in their free-form responses).

However, we identify 3 papers (~5%) that consider different aspects of understandability in their operationalisation. In [17] and [16], the authors consider understandability as a process that evolves over time across a single experiment on the legibility of multi-robot systems. To account for this, they use two prediction measures of understandability that a participant selects from the user interface (identifying the coordination objective from a set and the spatial goal of the multi-robot formation) and a performance measure of the time taken to provide an answer. Therefore, these three objective measures constitute the two dimensions of understandability - i) how accurate is the understanding of the goal and coordination objective and ii) how early the user understands. The experiment in [16] finds that both the independent variables, coordination objective and number of robots, have a statistically significant effect on the two

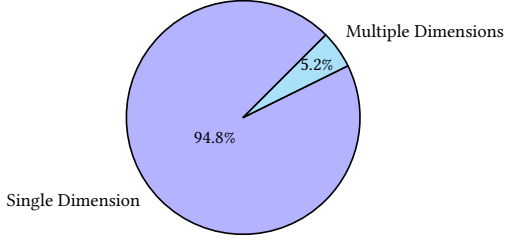


Fig. 8. Categorisation by single- or multi-dimensional view of understandability.

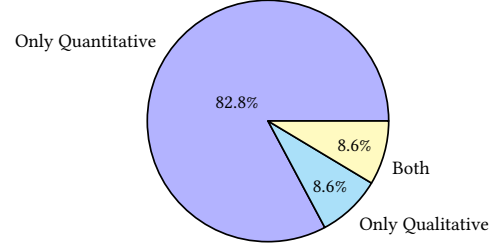


Fig. 9. Categorisation by the use of quantitative and qualitative measures.

dimensions. However, the experiment in [17] indicates that the different independent variables of the motion of the multi-robot system affect the two dimensions differently.

In [26], the authors distinguish between two types of understandability. The first is “*behaviour understanding*”, which in this context is measured by asking prediction-based questions about the navigation scenarios the robot encounters (e.g., whether the robot encountered a navigation problem and how this problem was detected). The second type is “*architecture understanding*”, which is measured by matching the functionality of the robot to its various physical components, such as its microphone, depth camera, and emergency button. The results indicate that while participants had a similar level of correctness for the *behaviour understanding* dimension, they had varying levels of correctness for the different components, thus having different levels of *architecture understanding*.

The analysis in this section highlights the lack of prior user studies that consider in any way the multi-dimensionality of understandability. Moreover, the distinct dimensions of understandability in the reviewed papers are only a subset of the theoretical taxonomies of multidimensional understandability (Sec. 2.2.2). For example, we find measurements in [16, 17] that refer to the *understanding what is the case* [6], but there are no measurements for *how/why is the case*. However, in this literature review, we aim to evaluate if there are works that acknowledge the multi-dimensionality of understanding in any form, since the research in multidimensional understanding is still in a very theoretical stage, as presented in Sec. 2.2.2.

4.3 Quantitative and Qualitative Measures

In this section, we categorise each paper based on its usage of quantitative or qualitative measures of understandability (see Table 2) based on the background literature from Sec. 2.2.3. While the distribution between objective and subjective evaluation methods is relatively balanced (Sec. 4.1), there is a pronounced preference for quantitative over qualitative measures. Specifically, approximately 91% of the eligible papers employ quantitative metrics. Among these, 48 papers (~83%) rely exclusively on quantitative evaluations, 5 papers (~9%) use only qualitative assessments, and another 5 papers (~9%) incorporate both types of measures (see Fig. 9).

For quantitative objective measures, a score is typically computed representing either a binary success or failure (e.g., whether a predictive measure is correct or not [72]) or the degree of success (e.g., how well the participant performs a task [88]). For subjective measures, the quantitative measure typically comes in the form of a questionnaire scale such as a Likert scale, as analysed in Sec. 4.1.

Similar to subjective measures, qualitative measures exhibit little to no standardisation, with each study presenting distinct questions to participants. Nonetheless, some common themes can be observed in how different works approach

Quantitative vs. Qualitative	Papers
Uses only quantitative measures	[9], [12], [16], [17], [19], [20], [21], [22], [26], [27], [28], [30], [35], [39], [41], [44], [45], [55], [60], [66], [67], [72], [74], [75], [76], [79], [78], [81], [84], [85], [88], [89], [90], [91], [93], [96], [98], [102], [108], [110], [111], [114], [116], [130], [131], [132], [133], [134]
Uses only qualitative measures	[31], [97], [115], [121], [124]
Uses both	[1], [38], [40], [113], [127]

Table 2. Categorisation of measures as either quantitative or qualitative.

Ecological Validity	Papers
Online study	[1], [12], [20], [22], [27], [28], [30], [38], [40], [41], [55], [66], [67], [74], [76], [81], [85], [88], [91], [93], [102], [110], [111], [114], [115], [116], [121], [124], [131], [134]
Laboratory study	[9], [16], [17], [19], [21], [26], [31], [35], [39], [44], [45], [60], [75], [79], [78], [84], [89], [90], [96], [97], [98], [108], [113], [127], [130], [132], [133]
"In the Wild" study	[72]

Table 3. Categorisation of the ecological validity of studies, conducted either online, in laboratory environments, or in the wild.

the qualitative evaluation of understandability. A common open-ended question is to ask participants to describe what they think the robot is doing/has done. For example, in [97], participants are asked "In your own words, what do you think the robot is doing?" after observing the robot for some period of time, while in [124], after watching a video of the robot, participants are asked what it did and what aspects of the behaviour were unexpected or significant, with special attention paid to actions that went beyond the commands of a human.

A significant portion of these papers asks participants to justify in their own words their response to some other question. For example, in [115], participants must justify why they selected a particular explanation as the "best" one, while in [40], participants are asked to justify their answer to a prediction question about the emotional state of a drone. In [113], participants are asked to justify their answers to subjective Likert-scale questions about their understanding and perception of their performance. In [38], participants are asked what they would like the robot to explain to them, and are also asked to justify their response.

In some papers, qualitative measures are employed to examine the participants' mental model of the robot. For example, in [31], participants are asked several questions to gauge their understanding of the robot, such as "How did the robot learn the positions of the objects?", while in [127], participants are asked to list 2-3 adjectives describing the emotional state of a robot. In other papers, participants are asked to introspect on their own understanding of the robot. For example, in [121], participants are shown videos of robots with expressive behaviours and are asked open-ended questions relating to the ease of understanding the robot and what factors influenced that process. In [1], participants are asked to explain the challenges they encounter in understanding a mission scenario as a drone operator.

4.4 Ecological Validity

For each of the eligible papers, we examine the context in which understandability is evaluated, categorising the studies based on whether they are conducted online, in a laboratory setting, or in-the-wild (see Table 3). This classification is informed by the discussion on ecological validity presented in Sec. 2.2.4.

Our analysis reveals that the majority of studies, 30 (~52%), are conducted online, typically carried out through crowdsourcing platforms such as Amazon Mechanical Turk and Prolific [112] (see Fig. 10). A similarly large proportion

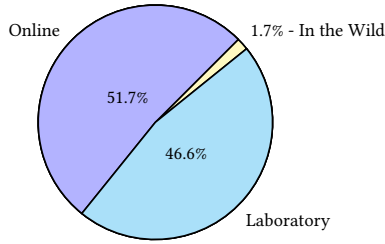


Fig. 10. Categorisation by study context.

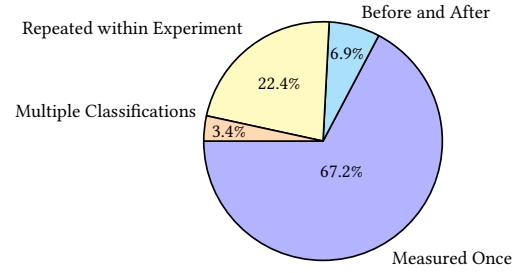


Fig. 11. Categorisation by measurement repetition.

Temporal Measurements	Papers
Measured once	[1], [9], [12], [17], [19], [20], [22], [27], [28], [30], [31], [39], [40], [41], [44], [45], [55], [60], [67], [72], [75], [76], [79], [78], [88], [89], [91], [96], [97], [108], [113], [114], [115], [124], [127], [130], [132], [133], [134]
Before and after	[26], [85], [93], [110], [111], [116]
Repeated within experiment	[16], [21], [35], [38], [66], [74], [81], [84], [85], [90], [93], [98], [102], [121], [131]
Longitudinal measurements	-

Table 4. Categorisation of the temporal aspect of the measurements of understandability. If a paper appears in multiple rows, it indicates that multiple measurements used in the paper are classified differently.

of studies, 27 (~47%), are conducted in laboratory settings. However, in some cases, participant interactions with the system are mediated entirely through a screen, minimising physical engagement with the experimental setup [108, 113, 130, 133]. This suggests that, despite participants being physically present in a laboratory, the distinction between online and laboratory settings is not always clear-cut, as the participant experience in some laboratory studies may be functionally equivalent to that of online studies.

In contrast to the prevalence of laboratory and online studies, only a single eligible paper reports an in-the-wild study [72]. In this work, the authors place a social robot unsupervised in a public space, where it elicits interactions from passers-by. Participants who engage with the robot have the opportunity to receive an explanation regarding the robot’s decision-making process, particularly its reasoning for engaging with individuals, e.g., why it waved at one person but not another. Understandability is then assessed using a prediction measure, that is, evaluating whether participants understand which participant the robot would engage with in a hypothetical situation.

4.5 Temporal Measurements

Following the discussion on the temporal aspect of measurements and novelty effects in Sec. 2.2.5, we further categorise the eligible papers to investigate whether researchers consider understandability as a construct that evolves over time. We initially developed four categories based on the frequency of measurements with respect to some explanatory intervention - *measured once*, *before and after*, *repeated within experiment*, and *longitudinal measurements*. However, as evident in Table 4, we find no user studies that conduct longitudinal measurements of understandability.

Our analysis reveals that the majority of the experiments follow the *measured once* approach, wherein understandability is typically assessed only after the completion of an experimental session that includes an explanatory intervention.

A total of 39 papers (~67%) contribute to this category (see Fig. 11). For example, several experiments measure subjective understanding using questionnaires after participants finish simulated navigation tasks [12, 19, 96, 114], or after watching videos of robot navigation [41].

The *before and after* category of papers collects the relevant measures before and after the explanatory intervention, usually to find changes in the measures after participants receive explanations. We find 6 papers (~10%) belonging to this category. For example, in [85], after watching a video of a robot perform a task, participants are asked to answer the question - “Do you think that something went wrong while the robot performed the task?”, repeating the same question before and after receiving an explanation. Similarly, in [110] and [111], participants are first asked to rate the understandability of a Pepper robot’s behaviour after watching videos with varied levels of surprise and desirability (e.g., Pepper providing entertainment in desirable and undesirable scenarios). Subsequently, the participants receive a verbal explanation for the behaviour and then rate how well this explanation was understandable.

The third category of the temporal measurements, *repeated within experiment*, consists of repeating the measurements multiple times during the course of a single experiment session. We find that 15 papers (~26%) fall into this category. The primary objective of taking multiple measurements is usually either to find differences in the measures across several tasks or to demonstrate the generalisation of the findings across multiple tasks. Notable examples include [81], where participants answer three types of questions a total of 30 times, corresponding to different traffic scenarios for an autonomous vehicle. In [66], understandability is measured each time after participants watch a video of a robot performing a task, which occurs three times. In [90], participants are shown 19 videos and have to evaluate changes in the environment that are detected by a robot and visualised using augmented reality. In [35], participants go through 4 iteration rounds over 5 tasks, where performance is assessed on each task.

Some articles are categorised into both *before and after* and *repeated within experiment* categories (marked “Multiple Classifications” in Fig 11) as they use more than one measure administered differently. For example, in [93], the researchers ask participants to identify the cause of robot failure after watching a video of the robot’s attempt but before receiving an explanation about the failure; then repeat the measure after they read an explanation. This *before and after* measurement continues for 7 trials, thus having multiple measurements in the same experimental session. The authors report a statistically significant difference in the failure cause prediction in *before and after* measures for groups that receive explanations, although for a control group (without explanations), no such differences were found. Similarly, in [85] the authors measure the participants’ ability to recognise a robot failure before and after providing explanations, repeating the measure three times for three different tasks and failure types. These articles contribute to ~3.4% of the total.

Although we hoped to find longitudinal studies that measure the effect of explanation on understandability, we have encountered none. In [38], the authors repeat the user study following identical procedures with a 15-month gap, although the objective of the replication study is to validate the robustness of the findings from the previous experiment. Even though the replication study finds similar results, it is not longitudinal since the authors purposefully exclude the participants from the initial study. Nevertheless, other longitudinal studies on related constructs, such as software usability, indicate that usability does not improve over time even after continuous interaction [52], but user frustration may decrease [77]. We may hypothesise that the understandability of a robot may improve over time with the addition of explanation in interaction. There are a few prior works in longitudinal XAI for non-embodied systems [59, 118, 129] that provide some possible trends, such as that users often adhere to initial beliefs despite explainability efforts [59, 118] and that additional information not necessarily leads to improved understanding [14, 118]. However, with the lack of longitudinal studies in the HRI field, the long-term effects of explainable HRI remain considerably unexplored.

Finally, we find interesting observations from a meta-analysis between the other review categories and the temporal aspect of measurements. Firstly, among the 19 studies that repeat measurements in any form (*before and after*, *repeated within experiment*), there is an equal share of experiments using only subjective and only objective measures, having 7 papers in each category (~37% for each), and 5 papers (~26%) incorporating both. Similarly, among the 39 studies that conduct only a single measurement, 15 (~38%) use only objective measures, 13 (~33%) use only subjective measures, and 11 (~28%) use both. However, ~58% of the studies that both conduct multiple measurements and use objective measures use performance-based measures, compared to the ~38% of papers using performance measures while making only a single measurement. Performance tasks are typically less intrusive because they do not require additional and explicit measurements beyond completing the task. For this reason, they may be preferred in studies that require multiple data points to avoid interrupting too many interactions.

5 Recommendations and Open Questions on the Evaluation of Understandability in HRI

After analysing the distribution and particularities of how previous studies in explainable HRI measure user understanding, we provide a set of general recommendations in the form of guidelines on how understandability evaluations should be carried out in explainable HRI user studies. Moreover, we explore several open questions facing the field of explainable HRI, which future research should investigate to provide a better comprehension and definition of the field.

5.1 The Understandability Construct

In our review, we have observed that a considerable share of the initially surveyed works state that their approach helps to improve user understanding through explainability, but then no measures related to understandability are provided. This is reflected in the *no measurement* exclusion criteria from Sec. 2, with a total of 17 papers excluded with this criteria. These results align with a recent review [103], which reveals that only 33% of user studies in explainable HRI actually measure understandability. Therefore, **we suggest that any study exploring the effects of explainability/transparency should measure understandability** (*recommendation*).

Legibility, interpretability, and even transparency and explainability are sometimes used almost as synonyms for understandability, although they all have specific definitions and interpretations. Those terms, including user understanding, can be used if they better fit the perspective of the study, but our suggestion is to always keep a link to the understandability construct. **We recommend using a common vocabulary to refer to user understanding. We suggest *understandability*** (*recommendation*), as it can be adopted for better searchability and differentiates from the word “understand”, which is commonly used with different meanings.

5.2 Objective and Subjective Measurements

Objective understandability tends to have a stronger impact on usability, ensuring that users can effectively interact with the robot to achieve a goal. In contrast, subjective understandability is more closely related to user trust, as an individual’s perception of their comprehension plays a crucial role in their confidence and comfort when interacting with robotic systems. **A thorough assessment should consider both objective and subjective understandability** (*recommendation*), as these two aspects influence different related factors. The literature on objective understandability measurement is highly diverse, largely because the approach depends on the specific task the robot performs. In many cases, prediction-based measures may be favoured, as these directly assess the user’s ability to anticipate and understand the robot’s behaviour. Performance-based measures can also be valuable, particularly in domains with shared tasks where human users rely on robotic support to achieve a common goal. Finally, feature relevance methods may be better

suiting for expert users who regularly use the system and should have more structured knowledge of how different input features affect the decision-making process. For subjective understandability, Likert scales are the most employed for quantitative assessments. In addition to standard Likert scales, open-ended questions can yield valuable qualitative insights. An effective approach might be to employ a think-aloud protocol in which participants verbalise their thoughts while completing the Likert scale, allowing researchers to capture additional context that complements the numerical responses.

However, **the field currently lacks widely accepted and validated questionnaires that can be systematically employed across different studies to ensure the comparability of results** (*open question*). A key challenge is determining whether a standardised approach is feasible across diverse explanation types or modalities and study settings, including both controlled laboratory experiments and real-world, in-the-wild deployments. Furthermore, longitudinal studies, which track changes in understandability over time, introduce additional complexities regarding consistency in questionnaire deployment and interpretation across different phases of user interaction. Future research should explore whether a universal framework for understandability evaluation can be established or whether study-specific adaptations will always be necessary. We believe that the creation of a validated standardised questionnaire for subjective understandability is possible, so we encourage future work to look into this problem. For objective understanding, we foresee that it can be dependent on domain, context, explanation modalities and types, and user profiles, but future research should clarify procedures or methodologies that can be followed depending on the circumstances.

Moreover, **the relationship between objective and subjective measures remains underexplored in the literature** (*open question*). Some works in explainable HRI have measured both objective and subjective understanding independently, without explicitly analysing their relation [30, 53, 66], while other works in the broader XAI field have acknowledged that “the relationship between subjective and objective understandability is an interesting topic for future work” [78]. A crucial issue arises when these measures exhibit a high level of disagreement. For instance, if users believe they understand a system well (high subjective understandability) but demonstrate poor performance in objective measures (low objective understandability), explanations would provide a placebo effect, generating higher trust in the system [29]. Nevertheless, with prolonged and repeated interactions, overconfidence might lead to inappropriate reliance on the system and, eventually, to frustration, decreased trust, and potential abandonment of the system. Conversely, if users have a high level of objective understandability but perceive their understandability to be low, they may exhibit unnecessary caution or disengagement. Future research should examine strategies to mitigate the discrepancies between objective and subjective understandability and explore the role of robotic communicative actions in shaping both of the measures.

5.3 Multi-dimensionality

We advocate for considering understandability as a multidimensional construct and suggest measuring different aspects of it (*recommendation*). As we have explored in this work, a significant gap between social sciences and technical explainable HRI studies remains open. One potential approach is to categorise robot behaviours and decision-making aspects based on their complexity and then assess user understanding within and across these categories. This approach enables a first multidimensional analysis of how different elements of the system contribute to overall understandability, although there are still many open questions regarding the multidimensionality of understandability.

Nevertheless, a significant challenge in assessing understandability is accounting for its inherently multi-dimensional nature. Understandability is not a monolithic construct; rather, it encompasses various facets which may manifest

differently depending on the context and user expectations. A critical open question is **how to effectively capture and measure these different dimensions within a single evaluation framework** (*open question*). Should these dimensions be assessed separately, or can a composite metric be developed that integrates them meaningfully? Additionally, future research should explore whether different dimensions of understandability contribute independently to overall user experience or whether they interact in ways that amplify or mitigate each other's effects. Moreover, the relationship of the multi-dimensional aspect with the objective and subjective understandability measures, including its subtypes, should be further analysed and formalised, to investigate if particular dimensions are measured more adequately by objective or subjective measures.

5.4 Quantitative and Qualitative Measurements

In evaluating understandability in human-robot interaction, it is essential to **employ both quantitative and qualitative measurement techniques** (*recommendation*). While many studies focus on quantitative measures, qualitative assessments can serve as a valuable complement, particularly in real-world, in-the-wild studies where collecting structured quantitative data may be challenging. The integration of both types of evaluation allows for a more comprehensive understanding of how users interpret and engage with robotic systems. An extensive qualitative evaluation might be infeasible due to a large number of participants or insufficient resources, but qualitative insights can be gathered during the initial phases of the study or to further analyse particular quantitative outcomes.

5.5 Longitudinal and into the Wild

As we have presented in Sec. 2.2.4, there have been concerns in the literature about the validity and replicability in real-world use cases of studies that are conducted in settings that force interactions with users (that is, online and laboratory settings). Moreover, as explored in Sec. 2.2.5, it is important to properly take into account the novelty effect and, if possible, to track the evolution of the measures through time. However, when it comes to measuring understandability, our review has revealed that only one study was conducted in the wild, while no studies have included the longitudinal component. Our recommendation regarding the study settings is to **aim for both longitudinal and in-the-wild studies** (*recommendation*), since they offer the most realistic impact on user understanding when assessing the effects of explainability. We encourage the HRI research community to value these studies while acknowledging the challenges that come with them, such as providing results that might be less statistically significant due to lower sample sizes and confounding variables arising from the unpredictability of the real world, yet with outcomes and trends that are going to be more relevant to shape the explainable HRI field.

5.6 Understandability Targets

It is also crucial to **clearly define the target level of understandability for a given study population** (*recommendation*), as previously suggested [32]. The goal should be for users to achieve 100% correctness within their designated target level, rather than requiring them to understand all aspects of the system. The necessary degree of understanding varies by user types; for instance, a lay user may require a more general understanding compared to a domain expert. In terms of objective measures, this means selecting appropriate tasks that align with the capabilities expected of the target users. For example, when using a performance measure, the target would be to successfully execute the task that the specific user type is required to perform in that use case. Subjective measures pose additional challenges, as individuals inherently interpret Likert scale boundaries in their own ways. To mitigate this variability, researchers could

provide concrete examples of what constitutes minimum and maximum understanding, helping to ground participants' responses and ensure more consistent interpretations.

Another unresolved issue concerns the precise definition of understandability targets for different user groups. While we have recommended that users should achieve a level of understanding appropriate to their role, the specific criteria for determining these targets remain ambiguous. **Future research should investigate how to systematically define understandability targets based on user needs, task demands, and contextual factors** (*open question*). A starting point can be the definitions of minimal and maximal understanding from the social sciences [6]. Moreover, it is essential to develop methodologies for validating whether users have achieved their intended level of understanding, while ensuring that the assessment process itself does not unduly influence user perceptions.

5.7 Link between Explanations, Understandability and Other Desiderata

A key question for future research is **how to effectively use the results of understandability evaluations to refine explanations provided by robots** (*open question*). Moreover, should explanations be personalised based on individual differences in cognitive abilities, prior knowledge, or familiarity with robotic systems? Another crucial aspect is how to adapt explanations to make users aware of their understanding, to keep a good balance between objective and subjective understandability, as we have already discussed in this section. Researchers should investigate methods for dynamically adjusting the granularity and format of explanations to optimise user comprehension while minimising unnecessary complexity and cognitive overload.

Finally, a broader challenge in explainable HRI research is **establishing a clear connection between the understandability construct and other key desiderata, such as trust, usability or performance** (*open question*), as illustrated in Fig. 1. While it is intuitively plausible that improved understandability enhances user trust and system usability, empirical studies are needed to quantify these relationships. Additionally, it remains unclear how different aspects of understandability —be they subjective, objective, or multi-dimensional— impact related factors. Future research should develop methodologies to systematically assess how changes in understandability influence user behaviour, robot perception, and overall system acceptance. By elucidating these connections, researchers can ensure that efforts to enhance understandability contribute meaningfully to broader improvements in human-robot interaction.

6 Conclusions

In this paper, we have devised a taxonomy to classify measures of understandability across five main aspects: the type of understandability measure, that is, objective or subjective; the dimensionality of the construct (unidimensional or multidimensional); whether they are measured quantitatively or qualitatively; the ecological validity of the experiments, classifying them into online, laboratory, or in-the-wild; and the temporal aspect of the measurements (single, before and after, repeated within experiment, or longitudinal). We then conducted a systematic review of the literature on the evaluation of explainable Human-Robot Interaction, focusing on the measurement of the understandability construct. We have identified 58 eligible papers that include user studies where understandability measures are included. We have categorised the papers based on our proposed taxonomy. Moreover, we have sub-categorised the objective measurements into prediction, performance, and feature selection tasks, while dividing subjective measurements between the ones that use new questionnaires, the ones employing confidence ratings, and the ones utilising or adapting existing questionnaires.

Our analysis reveals that the reviewed papers are unevenly distributed across the different categories, with a notable lack of longitudinal studies and in-the-wild experiments. We also find that most studies measure understandability as a

single, unidimensional construct, with only a few papers considering its multidimensional nature. Furthermore, we observe that the majority of studies employ quantitative measures.

We have provided a set of recommendations for researchers working in the field. We have highlighted the importance of using a common vocabulary and actually measuring understandability in explainability user studies. A further recommendation is to measure understandability both quantitatively and qualitatively, and to consider both objective and subjective measures in user studies that should aim to run longitudinally in the wild. In our work, we also reinforce the need to consider understandability as a multidimensional construct, and to define understandability targets for different user groups.

Finally, we have identified several open questions for further investigation, including the standardisation of questionnaires, the relationship between objective and subjective understandability, the multi-dimensionality of the construct, the definition of understandability targets, the methods to improve explanations based on understandability results, and the connection between understandability and other desiderata in HRI. We believe that addressing these questions will contribute to a deeper understanding of the role of understandability in HRI and will help to advance the field of explainable robotics.

Acknowledgments

This work has been supported by Horizon Europe Marie Skłodowska-Curie grant agreement No. 101072488 (TRAIL).

References

- [1] Ankit Agrawal and Jane Cleland-Huang. 2021. Explaining Autonomous Decisions in Swarms of Human-on-the-Loop Small Unmanned Aerial Systems. *AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (2021), 15–26.
- [2] Sule Anjomshoe, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1078–1088.
- [3] Ainhoa Apraiz, Ganix Lasa, and Maitane Mazmela. 2023. Evaluation of user experience in human–robot interaction: a systematic literature review. *International Journal of Social Robotics* 15, 2 (2023), 187–210.
- [4] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2021. Explaining in time: Meeting interactive standards of explanation for robotic systems. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–23.
- [5] Christoph Baumberger. 2014. Types of understanding: Their nature and their relation to knowledge. *Conceptus* 40, 98 (2014), 67–88.
- [6] Christoph Baumberger, Claus Beisbart, and Georg Brun. 2016. What is understanding? An overview of recent debates in epistemology and philosophy of science. *Explaining understanding* (2016), 1–34.
- [7] Paul Baxter, James Kennedy, Emmanuel Senft, Severin Lemaignan, and Tony Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th acm/ieee international conference on human-robot interaction (hri)*. IEEE, 391–398.
- [8] Tony Belpaeme. 2020. Advice to new human-robot interaction researchers. *Human-robot interaction: Evaluation methods and their standardization* (2020), 355–369.
- [9] Marijke Bergman, Sandra Bedaf, Goscha van Heel, and Janienke Sturm. 2020. Can I Just Pass by? Testing Design Principles for Industrial Transport Robots. In *CHIRA*. 178–187.
- [10] Meisam Booshehri, Hendrik Buschmeier, and Philipp Cimiano. 2024. Towards a BFO-based ontology of understanding in explanatory interactions. In *Proceedings of the 4th International Workshop on Data Meets Applied Ontologies in Explainable AI (DAO-XAI)*.
- [11] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59.
- [12] Martim Brandão, Amanda Coles, and Daniele Magazzeni. 2021. Explaining Path Plan Optimality: Fast Explanation Methods for Navigation Meshes Using Full and Incremental Inverse Optimization. *International Conference on Automated Planning and Scheduling* 31, 1 (2021), 56–64.
- [13] Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1–16.
- [14] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the ACM international conference on Intelligent User Interfaces*. 169–178.
- [15] Harriet R Cameron, Simon Castle-Green, Muhammad Chughtai, Liz Dowthwaite, Ayse Kucukyilmaz, Horia A Maior, Victor Ngo, Eike Schneiders, and Bernd C Stahl. 2024. A Taxonomy of Domestic Robot Failure Outcomes: Understanding the impact of failure on trustworthiness of domestic robots. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–14.

- [16] Beatrice Capelli, Maria Santos, and Lorenzo Sabattini. 2024. Towards the Legibility of Multi-robot Systems. *J. Hum.-Robot Interact.* 13, 2 (2024).
- [17] Beatrice Capelli, Valeria Villani, Cristian Secchi, and Lorenzo Sabattini. 2019. Understanding Multi-Robot Systems: on the Concept of Legibility. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 7355–7361.
- [18] J Adam Carter and Emma C Gordon. 2014. Objectual understanding and the value problem. *American Philosophical Quarterly* 51, 1 (2014), 1–13.
- [19] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan Explanations as Model Reconciliation – An Empirical Study. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 258–266.
- [20] Nandu Chandran Nair, Alessandra Rossi, and Silvia Rossi. 2024. Impact of Explanations on Transparency in HRI: A Study Using the HRIVST Metric. In *Social Robotics*. Springer Nature Singapore, 171–180.
- [21] Yuhang Che, Allison M. Okamura, and Dorsa Sadigh. 2020. Efficient and Trustworthy Social Navigation via Explicit and Implicit Robot–Human Communication. *IEEE Transactions on Robotics* 36, 3 (2020), 692–707.
- [22] Shenghui Chen, Kayla Boggess, and Lu Feng. 2020. Towards Transparent Robotic Planning via Contrastive Explanations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6593–6598.
- [23] Kerstin Dautenhahn. 2007. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems* 4, 1 (2007), 15.
- [24] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [25] Anca Dragan and Siddhartha Srinivasa. 2014. Familiarization to robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI '14)*. Association for Computing Machinery, New York, NY, USA, 366–373.
- [26] Leonie Dyck, Helen Beierling, Robin Helmer, and Anna-Lisa Vollmer. 2023. Technical Transparency for Robot Navigation Through AR Visualizations. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 720–724.
- [27] Matthias Eder, Clemens Könczöl, Julian Kienzl, Jochen A. Mosbacher, Bettina Kubicek, and Gerald Steinbauer-Wagner. 2024. Influence of Different Explanation Types on Robot-Related Human Factors in Robot Navigation Tasks. In *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 1084–1091.
- [28] Matthias Eder and Gerald Steinbauer-Wagner. 2024. Why Did My Robot Choose This Path? Explainable Path Planning for Off-Road Navigation. In *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 139–145.
- [29] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebo explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
- [30] Miguel Faria, Francisco S. Melo, and Ana Paiva. 2021. Understanding Robots: Making Robots More Legible in Multi-Party Interactions. In *IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. 1031–1036.
- [31] Helena Anna Frijns, Matthias Hirschmanner, Barbara Sienkiewicz, Peter Hönig, Bipin Indurkha, and Markus Vincze. 2024. Human-in-the-loop error detection in an object organization task with a social robot. *Frontiers in Robotics and AI* 11 (2024), 1356827.
- [32] Ferran Gebelli, Raquel Ros, Séverin Lemaignan, and Anaïs Garrell. 2024. Evaluating the Impact of Explainability on the Users’ Mental Models of Robots over Time. In *Late Breaking report in the IEEE International Conference on Robot and Human Interactive Communication*. IEEE Computer Society.
- [33] Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (2007), 619–619.
- [34] Stephen R Grimm. 2014. Understanding as knowledge of causes. In *Virtue epistemology naturalized: Bridges between virtue epistemology and philosophy of science*. Springer, 329–345.
- [35] André Groß, Amit Singh, Ngoc Chi Banh, Birte Richter, Ingrid Scharlau, Katharina J Rohlfing, and Britta Wrede. 2023. Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI* 10 (2023), 1236184.
- [36] Hatice Gunes, Frank Broz, Chris S Crawford, Astrid Rosenthal-von der Pütten, Megan Strait, and Laurel Riek. 2022. Reproducibility in human-robot interaction: Furthering the science of hri. *Current Robotics Reports* 3, 4 (2022), 281–292.
- [37] Amar Halilovic and Senka Krivic. 2024. Planning of explanations for robot navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5478–5484.
- [38] Zhao Han, Elizabeth Phillips, and Holly A. Yanco. 2021. The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself. *J. Hum.-Robot Interact.* 10, 4 (2021).
- [39] André Helgert, Lukas Erle, Sabrina Langer, Carolin Straßmann, and Sabrina C. Eimler. 2024. Towards Understandable Transparency in Human-Robot-Interactions in Public Spaces. In *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 1162–1169.
- [40] Viviane Herdel, Anastasia Kuzminykh, Andrea Hildebrandt, and Jessica R. Cauchard. 2021. Drone in Love: Emotional Perception of Facial Expressions on Flying Robots. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- [41] Nicholas J. Hetherington, Elizabeth A. Croft, and H.F. Machiel Van der Loos. 2021. Hey Robot, Which Way Are You Going? Nonverbal Motion Legibility Cues for Human-Robot Spatial Interaction. *IEEE Robotics and Automation Letters* 6, 3 (2021), 5010–5015.
- [42] Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2020), 1–31.
- [43] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.
- [44] Hidenori Itaya, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. 2024. Mask-Attention A3C: Visual Explanation of Action–State Value in Deep Reinforcement Learning. *IEEE Access* 12 (2024), 86553–86571.

- [45] Misbah Javaid and Vladimir Estivill-Castro. 2021. Explanations from a Robotic Partner Build Trust on the Robot's Decisions for Collaborative Human-Humanoid Interaction. *Robotics* 10, 1 (2021).
- [46] Kristiina Jokinen and Graham Wilcock. 2021. Do you remember me? Ethical issues in long-term social robot interactions. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 678–683.
- [47] Malte Jung and Pamela Hinds. 2018. Robots in the wild: A time for more robust theories of human-robot interaction. 5 pages.
- [48] Frank Keil. 2003. Categorisation, causation, and the limits of understanding. *Language and Cognitive Processes* 18, 5-6 (2003), 663–692.
- [49] Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57, 1 (2006), 227–254.
- [50] Matthias Kerzel, Jakob Ambsdorf, Dennis Becker, Wenhao Lu, Erik Strahl, Josua Spisak, Connor Gäde, Tom Weber, and Stefan Wermter. 2022. What's on Your Mind, NICO? XHRI: A Framework for eXplainable Human-Robot Interaction. *KI-Künstliche Intelligenz* 36, 3 (2022), 237–254.
- [51] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. 2023. Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 1829–1836. doi:10.1109/RO-MAN57019.2023.10309394
- [52] Jesper Kjeldskov, Mikael B Skov, and Jan Stage. 2010. A longitudinal study of usability in health care: Does time heal? *international journal of medical informatics* 79, 6 (2010), e135–e143.
- [53] Dimosthenis Kontogiorgos and Julie Shah. 2025. Questioning the Robot: Using Human Non-verbal Cues to Estimate the Need for Explanations. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*. 717–728.
- [54] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Congress of the International Ergonomics Association (IEA 2018)*. Springer International Publishing, Cham, 13–30.
- [55] Benjamin Krarup, Felix Lindner, Senka Krivic, and Derek Long. 2022. Understanding a Robot's Guiding Ethical Principles via Automatically Generated Explanations. In *IEEE 18th International Conference on Automation Science and Engineering (CASE)*. 627–632.
- [56] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.
- [57] Johannes Kraus. 2020. *Psychological processes in the formation and calibration of trust in automation*. Ph.D. Dissertation.
- [58] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.
- [59] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.
- [60] Shikhar Kumar, Yisrael Parmet, and Yael Edan. 2024. Exploratory user study on verbalization of explanations*. In *IEEE International Conference on Human-Machine Systems (ICHMS)*. 1–7.
- [61] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. 2018. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters* 3, 4 (2018), 4023–4030.
- [62] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 87–95.
- [63] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.
- [64] Thao Le, Tim Miller, Ronal Singh, and Liz Sonenberg. 2023. Explaining model confidence using counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11856–11864.
- [65] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5 (2013), 291–308.
- [66] Gregory LeMasurier, Alvika Gautam, Zhao Han, Jacob W. Crandall, and Holly A. Yanco. 2024. Reactive or Proactive? How Robots Should Explain Failures. In *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 413–422.
- [67] Gregory LeMasurier, Christian Tagliamonte, Jacob Breen, Daniel Maccalline, and Holly A. Yanco. 2024. Templated vs. Generative: Explaining Robot Failures. In *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 1346–1353.
- [68] James R Lewis. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78.
- [69] Christina Lichtenthäler and Alexandra Kirsch. 2016. Legibility of Robot Behavior : A Literature Review. (April 2016).
- [70] Peter Lipton. 2017. Inference to the best explanation. *A Companion to the Philosophy of Science* (2017), 184–193.
- [71] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [72] Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2024. What Would I Do If...? Promoting Understanding in HRI through Real-Time Explanations in the Wild. In *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 504–509.
- [73] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* (2025).
- [74] Christoforos Mavrogiannis, Patrícia Alves-Oliveira, Wil Thomason, and Ross A. Knepper. 2022. Social Momentum: Design and Evaluation of a Framework for Socially Competent Robot Navigation. *J. Hum.-Robot Interact.* 11, 2 (2022).

- [75] Heinrich Mellmann, Polina Arbuzova, Dimosthenis Kontogiorgos, Magdalena Yordanova, Jennifer X. Haensel, Verena V. Hafner, and Joanna J. Bryson. 2024. Effects of Transparency in Humanoid Robots - A Pilot Study. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 750–754.
- [76] Gaspar Isaac Melsion, Rebecca Stower, Katie Winkle, and Iolanda Leite. 2023. What’s at Stake? Robot explanations matter for high but not low-stake scenarios. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2421–2426.
- [77] Valerie Mendoza and David G Novick. 2005. Usability over time. In *Proceedings of the 23rd annual international conference on Design of communication: documenting & designing for pervasive information*. 151–158.
- [78] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. 2022. The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial Intelligence* 302 (2022), 103573.
- [79] Yazan Mualla, Igor Tchappi, Amro Najjar, Timotheus Kampik, Stéphane Galland, and Christophe Nicolle. 2020. Human-agent Explainability: An Experimental Case Study on the Filtering of Explanations. In *International Conference on Agents and Artificial Intelligence (ICAART): Special Session on Human-centric Applications of Multi-agent Technologies*. 378–385.
- [80] Mark A Neerinx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 15*. Springer, 204–214.
- [81] Daniel Omeiza, Konrad Kollnig, Helena Web, Marina Jirotko, and Lars Kunze. 2021. Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving. In *IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. 194–199.
- [82] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021).
- [83] Chung Hyuk Park, Raquel Ros, Sonya S Kwak, Chien-Ming Huang, and Séverin Lemaignan. 2020. Towards real world impacts: Design, development, and deployment of social robots in the wild. 600830 pages.
- [84] Ornnalin Phaijit, Claude Sammut, and Wafa Johal. 2023. User Interface Interventions for Improving Robot Learning from Demonstration. In *International Conference on Human-Agent Interaction*. Association for Computing Machinery, 152–161.
- [85] Pradip Pramanick, Luca Raggioli, Alessandra Rossi, and Silvia Rossi. 2024. Effects of Incoherence in Multimodal Explanations of Robot Failures. In *Companion of the International Conference on Multimodal Interaction*. Association for Computing Machinery, 6–10.
- [86] Harrison Preusse, Rebecca Skulsky, Marlena R Fraune, and Betsy Bender Stringam. 2021. Together we can figure it out: groups find hospitality robots easier to use and interact with them more than individuals. *Frontiers in Robotics and AI* 8 (2021), 730399.
- [87] Duncan Pritchard. 2014. Knowledge and understanding. In *Virtue epistemology naturalized: Bridges between virtue epistemology and philosophy of science*. Springer, 315–327.
- [88] David V. Pynadath, Nikolos Gurney, and Ning Wang. 2022. Explainable Reinforcement Learning in Human-Robot Teams: The Impact of Decision-Tree Explanations on Transparency. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 749–756.
- [89] Peizhu Qian and Vaibhav Vasant Unhelkar. 2024. Interactively Explaining Robot Policies to Humans in Integrated Virtual and Physical Training Environments. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 847–851.
- [90] Christopher Reardon, Jason M. Gregory, Kerstin S. Haring, Benjamin Dossett, Ori Miller, and Aniekan Inyang. 2024. Augmented Reality Visualization of Autonomous Mobile Robot Change Detection in Uninstrumented Environments. *J. Hum.-Robot Interact.* 13, 3 (2024).
- [91] Christopher Reardon, Kerstin Haring, Jason M. Gregory, and John G. Rogers. 2021. Evaluating Human Understanding of a Mixed Reality Interface for Autonomous Robot-Based Change Detection. In *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. 132–137.
- [92] Merle Reimann, Jesper van de Graaf, Nina van Gulik, Stephanie Van De Sanden, Tibert Verhagen, and Koen Hindriks. 2023. Social robots in the wild and the novelty effect. In *International Conference on Social Robotics*. Springer, 38–48.
- [93] David A. Robb, Xingkun Liu, and Helen Hastie. 2023. Explanation Styles for Trustworthy Autonomous Systems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Association for Computing Machinery, 2298–2300.
- [94] David A Robb, Francisco J Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 384–392.
- [95] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* (2023).
- [96] Stephanie Rosenthal, Peerat Vichivanives, and Elizabeth Carter. 2022. The Impact of Route Descriptions on Human Expectations for Robot Navigation. *J. Hum.-Robot Interact.* 11, 4 (2022).
- [97] Alexandros Rotsidis, Andreas Theodorou, Joanna J. Bryson, and Robert H. Wortham. 2019. Improving Robot Transparency: An Investigation With Mobile Augmented Reality. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1–8.
- [98] Karina A. Roundtree, Jason R. Cody, Jennifer Leaf, H. Onan Demirel, and Julie A. Adams. 2022. Transparency’s Influence on Human-collective Interactions. *J. Hum.-Robot Interact.* 11, 2 (2022).

- [99] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.
- [100] Tatsuya Sakai and Takayuki Nagai. 2022. Explainable autonomous robots: a survey and perspective. *Advanced Robotics* 36, 5–6 (2022), 219–238.
- [101] Tamie Salter, François Michaud, and Hélène Larouche. 2010. How wild is wild? A taxonomy to characterize the ‘wildness’ of child-robot interaction. *International Journal of Social Robotics* 2 (2010), 405–415.
- [102] Lindsay Sanneman and Julie A. Shah. 2022. An Empirical Study of Reward Explanations With Human-Robot Interaction Applications. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8956–8963.
- [103] Svenja Y Schött, Rifat Mehreen Amin, and Andreas Butz. 2023. A literature survey of how to convey transparency in co-located human–robot interaction. *Multimodal Technologies and Interaction* 7, 3 (2023), 25.
- [104] Sonali K Shah and Kevin G Corley. 2006. Building better theory by bridging the quantitative–qualitative divide. *Journal of management studies* 43, 8 (2006), 1821–1835.
- [105] Catharina Vesterager Smedegaard. 2019. Reframing the role of novelty within social HRI: from noise to information. In *2019 14th acm/ieee international conference on human-robot interaction (hri)*. IEEE, 411–420.
- [106] Barry Smith. 2013. Classifying processes: an essay in applied ontology. *Classifying Reality* (2013), 101–126.
- [107] David Sobrin-Hidalgo, Ángel Manuel Guerrero-Higueras, and Vicente Matellán-Olivera. 2025. Generating Explanations for Autonomous Robots: a Systematic Review. *IEEE Access* (2025).
- [108] Utkarsh Soni, Sarath Sreedharan, and Subbarao Kambhampati. 2021. Not all users are the same: Providing personalized explanations for sequential decision making problems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6240–6247.
- [109] Timo Speith and Markus Langer. 2023. A new perspective on evaluation methods for explainable artificial intelligence (XAI). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, 325–331.
- [110] Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot’s Self-Explanations on How Humans Understand and Evaluate Its Behavior. In *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 619–627.
- [111] Sonja Stange and Stefan Kopp. 2021. Effects of Referring to Robot vs. User Needs in Self-Explanations of Undesirable Robot Behavior. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 271–275.
- [112] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences* 21, 10 (2017), 736–748.
- [113] Oliver Struckmeier, Mattia Racca, and Ville Kyrki. 2019. Autonomous Generation of Robust and Focused Explanations for Robot Policies. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1–8.
- [114] Roykrong Sukkerd, Reid Simmons, and David Garlan. 2020. Tradeoff-Focused Contrastive Explanation for MDP Planning. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1041–1048.
- [115] Zachary Taschdjian. 2020. Why Did the Robot Cross the Road? A User Study of Explanation in Human-Robot Interaction. In *HCI International - Late Breaking Papers: Multimodality and Intelligence*. Springer International Publishing, 527–537.
- [116] Lena Trigg, Brandon Morgan, Alexander Stringer, Lacey Schley, and Dean F. Hougen. 2024. Natural Language Explanation for Autonomous Navigation. In *AIAA DATC/IEEE Digital Avionics Systems Conference (DASC)*. 1–9.
- [117] John Dewain Trout. 2007. The psychology of scientific explanation. *Philosophy Compass* 2, 3 (2007), 564–591.
- [118] Joe Tullio, Anind K Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 31–40.
- [119] Jasper van der Waa, Elisabeth Nieuwborg, Anita Cremers, and Mark Neerinx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.
- [120] Sophie van der Woerd and Pim Haselager. 2016. Lack of effort or lack of ability? robot failures and human perception of agency and responsibility. In *Benelux conference on artificial intelligence*. Springer, 155–168.
- [121] Marieke Van Otterdijk, Diana Saplaan Lindblom, Bruno Laeng, and Jim Torresen. 2024. An Exploratory Study on People’s Intuitive Understanding of Expressive Robot Behavior. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 1072–1076.
- [122] Louise Veling and Conor McGinn. 2021. Qualitative research in HRI: A review and taxonomy. *International Journal of Social Robotics* 13 (2021), 1689–1709.
- [123] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. 2023. A Survey of Evaluation Methods and Metrics for Explanations in Human–Robot Interaction (HRI). In *ICRA2023 Workshop on Explainable Robotics*.
- [124] Nick Walker, Kevin Weatherwax, Julian Allchin, Leila Takayama, and Maya Cakmak. 2020. Human Perceptions of a Curious Robot that Performs Off-Task Actions. In *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 529–538.
- [125] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–24.
- [126] Alan F. T. Winfield, Serena Booth, Louise A. Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I. Mutttram, Joanna I. Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark A. Underwood, Robert H. Wortham, and Eleanor Watson. 2021. IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and AI* 8 (2021).

- [127] Hannen Wolfe, Yiheng Su, and Jue Wang. 2024. Dimensional Design of Emotive Sounds for Robots. In *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 791–799.
- [128] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings 18*. Springer, 274–289.
- [129] Rayoung Yang and Mark W Newman. 2013. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the ACM international joint conference on Pervasive and ubiquitous computing*. 93–102.
- [130] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science Robotics* 7, 68 (2022), eabm4183.
- [131] Mehrdad Zakershahra, Ze Gong, Nikhilesh Sadassivam, and Yu Zhang. 2020. Online Explanation Generation for Planning Tasks in Human-Robot Teaming. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6304–6310.
- [132] Wenjuan Zhang, David Feltner, David Kaber, and James Shirley. 2021. Utility of functional transparency and usability in UAV supervisory control interface design. *International Journal of Social Robotics* 13, 7 (2021), 1761–1776.
- [133] Xuan Zhao, Tingxiang Fan, Dawei Wang, Zhe Hu, Tao Han, and Jia Pan. 2020. An Actor-Critic Approach for Legible Robot Motion Planner. In *IEEE International Conference on Robotics and Automation (ICRA)*. 5949–5955.
- [134] Mengyu Zhong, Marc Fraile, Ginevra Castellano, and Katie Winkle. 2023. A case study in designing trustworthy interactions: implications for socially assistive robotics. *Frontiers in Computer Science* 5 (2023), 1152532.
- [135] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology* 111, 4 (2016), 493.